

# Identifying Counterfeit Medicines

## BTEP Project Report

*Submitted by:*

**Malav Doshi (AU1940017)**

in partial fulfillment for the award of the degree

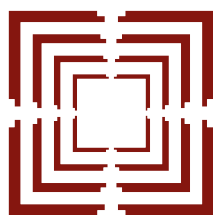
of

**BACHELOR OF TECHNOLOGY**

in

**Computer Science and Engineering**

at



**Ahmedabad**  
University

**School of Engineering and Applied Science (SEAS)**

**Ahmedabad, Gujarat**

**May, 2023**

## DECLARATION

I hereby declare that the project entitled “**Identifying Counterfeit Medicines**” submitted for the B. Tech. (**Computer Science and Engineering**) degree is my original work and the project has not formed the basis for the award of any other degree, diploma, fellowship or any other similar titles.

**Signature of Student**

**Date:**

**Place:**

## CERTIFICATE

This is to certify that the project titled “**Identifying Counterfeit Medicines**” is the bona fide work carried out by **Malav Doshi**, a student of B. Tech. (**Computer Science and Engineering**) of School of Engineering and Applied Science at Ahmedabad University during the academic year 2022-2023, in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in **Computer Science and Engineering** and that the project has not formed the basis for the award previously of any other degree, diploma, fellowship or any other similar title.

This project under the supervision of the industry mentor **Dr. Ashish Kundu**.

Signature of Industry Mentor

Signature of Faculty Mentor

Date:

Date:

Place:

Place:

## ACKNOWLEDGEMENT

I would like to acknowledge and express my sincere gratitude to Dr. Ashish Kundu and Professor Amit Nanavati for their invaluable guidance, support, and assistance throughout the course of this project. Despite encountering several challenges and bumps along the way, their unwavering commitment and expertise have been instrumental in steering this project toward successful completion. I am fortunate to have had the opportunity to work under their supervision and benefit from their wisdom and experience.

I would also like to extend my gratitude to the entire faculty and staff of Ahmedabad University for providing a conducive environment and resources that have facilitated the successful execution of this project.

## **Abstract**

Identifying Counterfeit medicines using concepts of Machine Learning, NLP, and Blockchain is a project that aims to identify fake medicines via methods of feature extraction of the medicines and using cross-analysis to compare the metadata. 20% Indian Markets have fake medicines that can adversely affect patients. According to the World Health Organization (WHO), up to two billion people worldwide lack access to necessary medicines, vaccines, medical devices including in vitro diagnostics, and other health products, which creates a vacuum that is too often filled by substandard and falsified products. India remains the leading provenance economy of counterfeit pharmaceuticals, originating 53% of the total seized value of counterfeit pharmaceutical products and medicines worldwide in 2016. The outcome of this project is to develop an interface that can verify the authentication of the drug.

# Table of Contents

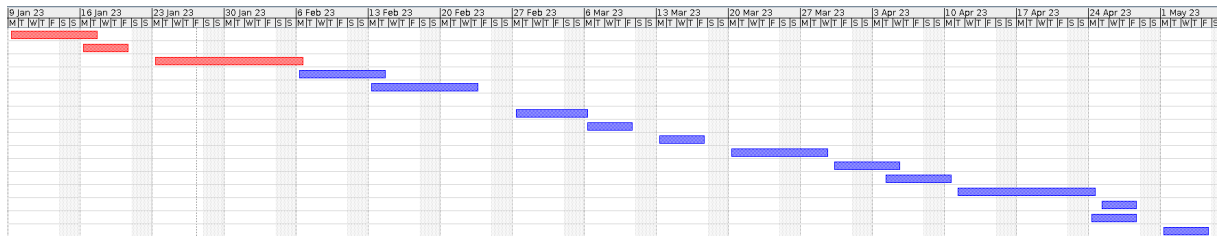
Declaration . . . . .	i
Certificate . . . . .	ii
Acknowledgment . . . . .	iii
Abstract . . . . .	iv
Table of Contents . . . . .	v
<b>List of Figures</b>	<b>vi</b>
<b>Gantt Chart</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Project Definition . . . . .	2
1.2 Project Objectives . . . . .	2
1.3 Project Deliverables . . . . .	2
1.4 Specifications . . . . .	3
<b>2 Literature Survey</b>	<b>4</b>
2.1 Related Work . . . . .	5
<b>3 Methodology</b>	<b>7</b>
<b>4 Results</b>	<b>11</b>
4.1 Project Outcomes . . . . .	13
4.2 My Contributions to the Project . . . . .	13
4.3 Learning Outcomes . . . . .	14
4.4 Research Outcomes . . . . .	14
4.5 Real World Applications . . . . .	15
<b>5 Conclusion</b>	<b>17</b>
<b>Bibliography</b>	<b>17</b>
<b>Appendices</b>	<b>19</b>

# List of Figures

2.0.1 Symbol Identification[1]	4
2.0.2 Results of Blister Identification - Front and Back[2]	5
2.1.1 Results from XRF Minipal2 [3]	5
2.1.2 Spectroscopy system design[4]	6
3.0.1 Sample Low-quality image scraped via Beautiful-Soup Scraping method	7
3.0.2 General Architecture of the System	10
4.0.1 Cosine Similarity Results for Identification	12
4.0.2 Consistency Analysis	12
4.0.3 Consistency Analysis	12

# Gantt Chart

	Name
1	Literature Survey
2	Dataset for List (identifying sources)
3	Scraping Using multiple methods
4	Vecotorising dataset
5	Processing Dataset and Vecotrising again
6	Hosting on Digital Ocean
7	Learning flutter
8	Reevaluaton Dataset & Scraping from different source
9	ML model test
10	Image processing (cleaning, augmentation)
11	Exploring ways with dataset
12	Feature Extraction of Images
13	Processing Text (Google Vision API)
14	Storing features in JSON (to IPFS)
15	JSON Comparison and grpah evaluation
16	Project Report





# Chapter 1

## Introduction

Fake medicines are a major problem in India and around the world. According to a paper by industry body ASSOCHAM, fake medicines constitute nearly one-third of all drugs sold in NCR and US \$ 4.25 billion of the total US\$ 14-17 billion of the domestic drugs market. Another report by the US Trade Representative Office on intellectual property rights protection in 2019 stated that up to 20% of the medicines sold in India were counterfeit.

Globally, the World Health Organisation estimates that up to 1% of medicines available in the developed countries are likely to be counterfeit. This figure rises to 10% globally, although in some developing countries it is 50%. The WHO also estimated that the share of counterfeit medicines on the market ranges from over 10% of total sales in low and middle-income countries to 1% in developed countries.

The problem of fake medicines is not limited to developing countries. In 2018, Pfizer identified 95 fake products in 113 countries, up from 29 fakes in 75 countries in 2008. The growth of e-commerce has also contributed to this trend by making it easier to purchase medicines online, often from unauthorized sources.

Fake medicines can have serious consequences for patients. They may contain no active ingredient, the wrong active ingredient, or the wrong amount of the correct active ingredient. They are also found to commonly contain corn starch, potato starch, or chalk. Some substandard and falsified medical products have been toxic in nature with either fatal levels of the wrong active ingredient or other toxic chemicals. Substandard and falsified medical products are often produced in very poor and unhygienic conditions by unqualified personnel and contain unknown impurities, and are sometimes contaminated with bacteria.

The project aims at identifying counterfeit medicine using concepts of NLP and Machine Learning. This project aimed to identify ways to detect features of fake medicines that can be used to distinguish between real and fake medicines. The project involves three phases. The first phase is the collection of the dataset. This is done by initially extracting the list of images followed by scraping. The second phase extracting various features of the medicines over a large dataset. This is to identify the observed features of medicines. The last step involves storing the metadata of over IPFS immutable chain post using a digital signature.

The project focuses on using cross-analysis due to the nature of the dataset obtained. The scarcity of data makes it impossible to use models such as deep learning, denseness, etc for classification. However, the process of the project justifies the decisions taken throughout.

**Key Words:** NLP, Feature Extraction, Scraping, Digital Signature, IPFS

## 1.1 | Project Definition

The problem statement for this project is to address the issue of counterfeit medicines, which is a serious global public health problem. Counterfeit medicines are fake or substandard medications that are sold as genuine drugs. They can contain the wrong amount of active ingredients, no active ingredients at all, or harmful substances that can cause serious harm or even death.

Identifying counterfeit medicines is a machine that includes using various technologies to collect, process, and explore profound ways to go about the classification of these images. The project uses a combination of computer vision, natural language processing and cross-analysis.

## 1.2 | Project Objectives

The aim of this project is to develop a system that can identify counterfeit medicines using natural language processing (NLP) and machine learning (ML) techniques. The system collects a list of medicines using web scraping techniques and obtains images of the medicines from various sources. Then, it uses the and other techniques to identify features of the images and stores these features as metadata of each medicine. The metadata is encrypted using the RSA algorithm and stored on the IPFS file system to ensure security and integrity.

Overall, the proposed system aims to provide a minimum viable and reliable method for identifying counterfeit medicines, which can help protect individuals and healthcare systems from the harmful effects of fake drugs.

## 1.3 | Project Deliverables

1. **Data scraping script:** A script that can scrape data from various sources and extract a list of medicines for analysis.
2. **Image processing module:** A module that can process images of medicines using the cloud vision api and other techniques to identify features and extract metadata.
3. **Metadata storage and encryption system:** A system that can store the metadata associated with each medicine in a secure and encrypted format using the RSA algorithm and IPFS file system.
4. **NLP analysis module:** A module that can analyze the metadata associated with each medicine using NLP techniques to identify patterns that may indicate the presence of counterfeit medicines.

5. **Test results:** Comprehensive test results to evaluate the accuracy and effectiveness of the system in identifying counterfeit medicines.

## 1.4 | Specifications

Please note the following parameters present on which the project was built upon.

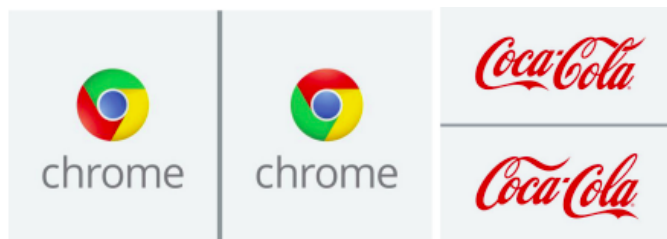
1. **Operating System:** This project was built on Ubuntu 20.04 LTS. It is recommended to use this operating system. However, it should be consistent with other Operating Systems as well.
2. **Programming Language:** This project was mainly developed on Python 3.8.10 and python libraries.
3. **Libraries:** The project requires image processing techniques to extract metadata from the images of medicines. Libraries such as OpenCV, Pillow, and Scikit-image have been used for this purpose.
4. **Scraping:** For scraping, digital ocean server of 8 GB memory/4 AMD vCPUs/160 GB Disk/ Ubuntu 20.04(LTS) x64 was used.
5. **IPFS File System:** The project requires the IPFS file system to store the encrypted metadata. The IPFS daemon can be installed on the local system or hosted on a remote server.

# Chapter 2

## Literature Survey

The problem of counterfeiting medicines has persisted over a long period of time, spanning through past and various geographies, and continues to pose a significant threat to public health and safety. Several discussions have been taken to using machine learning however the implementation of these discussions is at an early stage.

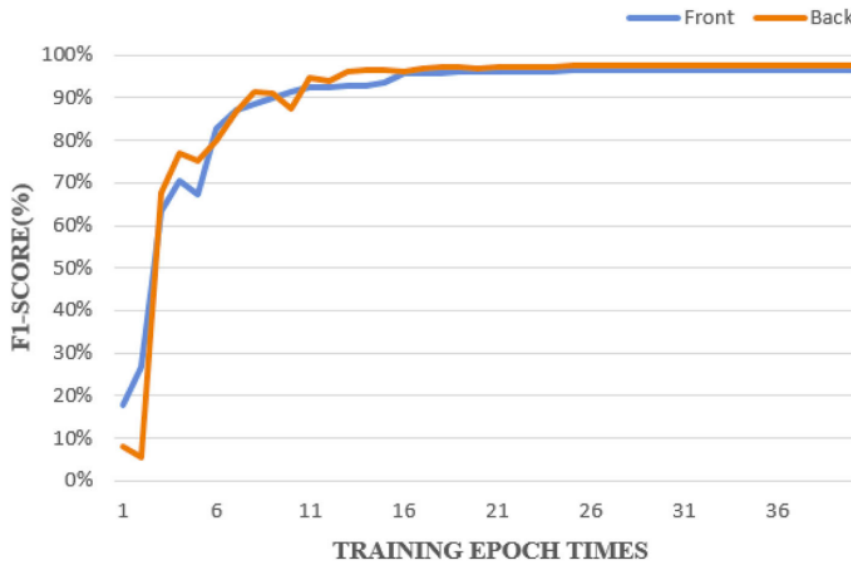
[1] proposes a non-invasive alert system that tries to distinguish the image of the logo of given medicine from the images of already approved images using methods of Deep learning (As in figure 2.0.1). If it does not identify from the given approved list of logos, it sends an alert of the medicine is fake. This uses a pre-trained VGG-16 -a CNN with 16 layers that have weights Lea model through transfer learning and obtained 96% accuracy for brand recognition and 84% for test accuracy in fake detection.



**Figure 2.0.1:** Symbol Identification[1]

Meanwhile, [5] proposes a system for the management of drug supply and recommends drugs for the pharmaceutical industry. It addresses counterfeit medicines by developing drug tracking and management through a blockchain-based platform that enables stakeholders to verify the authenticity and provenance of drugs. In addition, it builds a system to analyze patient data, such as medical history and genetic information to identify drugs that are most likely to be effective and safe for patients. This way the patient can get a recommendation/prescription.

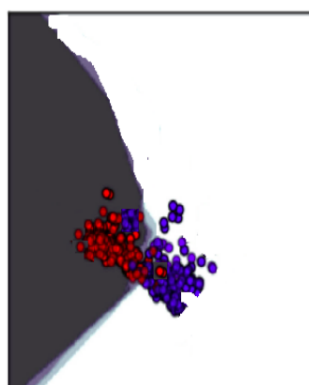
[2] performs drug identification (Taiwan-specific region using Yolo). The discussion revolves around solving Look alike and Sound alike (LASA) medical errors. This implementation focuses on the identification of blister-specific images. The accuracy of the training and validation of over 250 packages led to an accuracy greater than 90%. The goal of the paper was to train over look-alike medicines that can reduce subsequent human error and can be implemented in real-life. The results can be seen in 2.0.2.



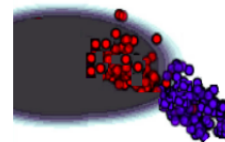
**Figure 2.0.2:** Results of Blister Identification - Front and Back[2]

## 2.1 | Related Work

In this section, we discuss a system that has a completely novel approach to solving the problem. For instance, [3] proposes XRF-Minipal2 which considers the case of extreme forgery to a level that features extraction or supply chain management cannot solve counterfeiting. The paper develops an approach over 10 samples of Tenormin tablets from different manufacturers and performs heavy chemical analysis to detect counterfeit Tenermin. It uses KNN and SVM methodology to perform the same. The Machine learning model is built on the post-analysis data of X-ray fluorescence that can perform procedures to identify the material's elemental composition. Using this technique, the system was able to identify the active component of 7 real medicines which was not present on the 3 counterfeit ones.



(a) Results from KNN



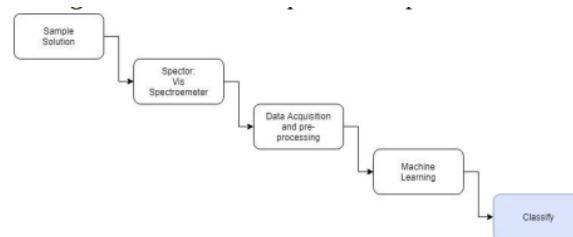
(b) Results from SVM

**Figure 2.1.1:** Results from XRF Minipal2 [3]

Another viable approach is discussed in [6] where the detection of counterfeit medicines is done using Hyperspectral Sensing. The near-infrared hyperspectral device is used to capture the spectral signature of medicines. The fake medicine components were included by adulterating powdered medicines. Spectral signatures were captured and machine

learning was used to achieve the detection of the ones with adulteration. The result showed more than 90% classification accuracy. A multi-layered perceptron was used to create a neural network and classification.

[4] uses the low-cost spectroscopy-based system to detect counterfeit medicines and adulterated food. Spectroscopy which is often used for detecting contaminants is expensive. In this case, the visible spectrometer is used to capture data which is preprocessed and identified via a model with CNN, SVM, and Logistic Regression models implemented. All models showed accuracy above 90%. Figure 2.1.2 shows the system design of the implementation method.



**Figure 2.1.2:** Spectroscopy system design[4]

# Chapter 3

## Methodology

The methodology to approach the problem has changed its shape in several aspects. The lack of a publicly available dataset and lack of machine learning accuracies and techniques to deal with such a noisy and considerably small dataset. The initial dataset consists of a list of medicinal strips that were scraped. This dataset contained more than 173,000 entries. On observing trends, there was no specific website that contained more than 2-3 images.

Following this, I scraped the images from Google search. This involved a lot of complications and hence had to explore several methods. These methods are:

- BeautifulSoup: This is a Python library that parses through HTML and XML documents. It is useful in going through structured and unstructured HTML. Now, upon using this I was unable to parse post the thumbnail images which led to a collection of low-quality images.



**Figure 3.0.1:** Sample Low-quality image scraped via Beautiful-Soup Scraping method

- Google scraping API: It provides a simple API for accessing Google search results and can be used to extract a range of data, including URLs, snippets, and metadata. Upon using this, it had a limited number of requests available and a maximum limit of 10k requests per day limit (with a premium subscription) which in our case can take about 6 months for data collection.
- Selenium and Chromium Webdriver: Selenium is used for browser automation. The Selenium API uses the WebDriver protocol to control web browsers like Chrome, Firefox, or Safari. Selenium can control both, a locally installed browser instance, as well as one running on a remote machine over the network. Here I have implemented 'headless' which runs Chrome in the background

The image collection had a limit to go up to 5 images which makes the total number of images go up to 869,045 images at maximum which potentially holds a large number of storage that cannot be stored nor processed locally. This led to the need for hosting and setting up a server on digital ocean with high-end specifications. The code was modified to use multi-threading with about 10 workers. However, the scraping of data failed several times due to throttle of CPUs or sqlite reaching maximum number of requests(to maintain progress count). This was solved by adjusting number of workers and using PostgreSQL (Used to maintain large amounts of dataset).

In parallel, I started analyzing a dataset of images of 3000 medicines out of a total of 173,000. The first step was to vectorize pairwise to identify a base image for character recognition. Upon vectorizing, the images showed results that are inconsistent and of poor quality, and do not provide any useful or meaningful insights. After processing images, the results did not improve. On further manual analysis, it could be observed that the images often had issues such as watermark, poor quality, include other objects, or having completely unrelated images from the search query. This led to a re-evaluation of this particular dataset and looking for alternatives. This dataset because of being inconsistent could not be further used for feature extraction. Note: This dataset which might not be useful in this use-case might be useful in cases in which the user has access a larger amount of meta information leading to better search queries or processing this dataset making it useful.

Further, started scraping images from a single source -Tata 1mg website. This website though consisted of a limited number of images but was consistent and could be used for better feature extraction of images. The json consisting of <image data> of images were stored in Digital ocean spaces.

The next step was trying for classification to identify the medicine type. Using multi-class classification using Yolo-v5 on the collected dataset led to a validation accuracy of 99.1%. A total of 110 classes of medicines each containing up to 5 images were considered. However, the classification rate on unseen data was nonexisting and completely random. Further, upon augmentation and trying several iterations and methodologies, the classification of medicine did not yield the desired outcome due to the inconsistencies and lack of dataset. An alternative method of approach was feature extraction of each image extensively. Now various features such as:

- Colour Moments
- Text recognition
- Text optimization and individual identification
- Text coordinates
- Shape features
  - Mean Area
  - Std Area
  - Mean Perimeter



- ☐ Std Perimeter
- ☐ Mean Aspect Ratio
- ☐ Mean Aspect Ratio
- ☐ Mean Centroid X
- ☐ Std Centroid X
- ☐ Mean Centroid Y
- ☐ Std Centroid Y
- Texture features
  - ☐ Contrast
  - ☐ Dissimilarity
  - ☐ Homogeneity
  - ☐ Energy
  - ☐ Correlation
- Pattern Features
- Entropy
- Vector

These features were considered to find a common meta-information with unique hyper-tuned parameters creating unique values for medicines and building cosine similarity over them. These features were observed by building a local dataset with physical medicines and accordingly optimizing the text.

Following is the list of variables and their denotations:

- Text list:  $d_t$
- Colour features:  $d_c$
- Texture features:  $d_{tx}$
- Shape features:  $d_s$
- Pattern features:  $d_p$

**Note:** Text coordinates were not considered because of its inability to draw precise rectangles when fed to the annotation function.

Now storing these features in JSON and comparing these features -Jaccardian Distance and Euclidian Distances are used. Now following is the methodology to obtain a common score from the information obtained.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (3.0.1)$$

Now we consider the following for it to be uniform amongst the others i.e we observe that when the distance increases the similarity between the metadata between JSON files decreases. Hence our parameter  $d_t$  should behave in similar fashion.

$$d_t = 1 - J(A, B) \quad (3.0.2)$$

Similarly the euclidian distance for  $d_c, d_{tx}, d_s, d_p$  is calculated as:

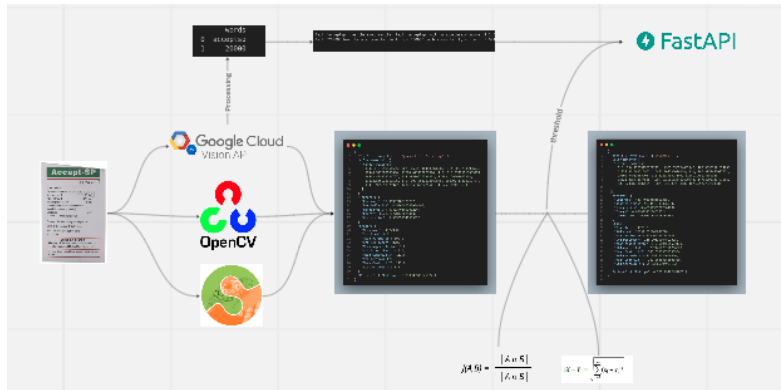
$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad (3.0.3)$$

Using the above equation we calculate a 20 dimensional for colour moments, 5 dimensions of texture features, 10 dimensions of shape features, 1 dimension of pattern features.

Now we consider the following equation to compute the absolute value of the distance between the following two parameters:

$$FinalMetric = \sqrt{d_t^2 + d_c^2 + d_{tx}^2 + d_s^2 + d_p^2} \quad (3.0.4)$$

Following this we get a maximum of  $\binom{12}{2}$  which equals to 66 values.



**Figure 3.0.2:** General Architecture of the System

# Chapter 4

## Results

Studies have showcased various methods for identifying counterfeit medicines, leveraging data from centralized sources such as logos or identifying information through advanced techniques like spectroscopy and hypersensing (as discussed in Chapter 2). However, the different approaches have different issues with them. Identifying logos might not be able to deal with new or unrecognized legitimate logos and might generate false alerts. While methods such as spectroscopy or hypersensing might face an issue when a contaminant goes undetectable. These systems might not be scalable due to their reliance on the existing dataset and inability to include the new versions in the machine learning models.

In one such instance, a dataset with a high validation rate of 99.1% was unable to support multi-class classification due to the overfitting of the model. Even after applying data augmentation techniques such as blur, Gaussian noise, and rotation, one class had only 12 images, leading to the overfitting of the test dataset.

To address this issue, an alternative approach was explored. One option was to use Generative Neural Networks, commonly used for generating "deep-fake" or "this person does not exist" images, to create fake images of the medicine. However, GANs require thousands of images to generate convincing fakes. Another approach involved identifying common features of the medicine images, such as shape, pattern, and texture, using feature extraction techniques. However, this approach posed its own challenges, such as the inability of OCR modules like `pytesseract` to accurately extract text coordinates.

To overcome these challenges, a Cloud Vision API was employed, which was pre-trained on a large dataset and proved effective in identifying the intended text. The feature extraction approach was used to capture shape, pattern, and texture information, as outlined in Chapter 3. This approach is being proposed as an outcome of the present project which uses cross-analysis to iterate over the already stored meta information of medicinal class to find the confidence level of the similarity found.

Now one of the obtained results is processing through the dataset and being able to get similarity using `Sequence Matcher`. Now following is the result when cosine similarity was used. We can see in figure 4.0.1 that indicates that the cosine similarity is unable to identify. The reason might be that the vectors of the word list are sparse and hence not able to capture relevant information or correlated dimensions might have caused cosine similarity to be similar with false data (it does not take covariance between dimensions

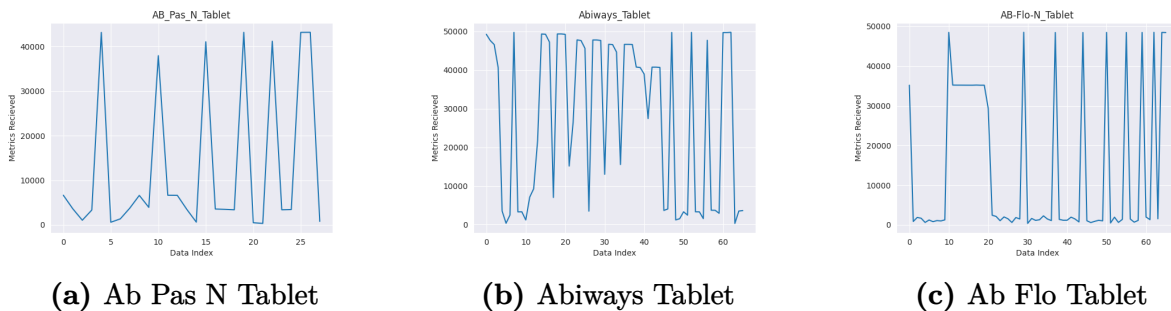
into account). However, the sequence matcher is able to find a better similarity factor as it looks for the Longest common subsequence (LCS). Here it works better than cosine similarity in this situation because here order and position of words are important factors in determining similarity.

	db_names	acceptsp
84348	larazolemd kid tablet orange	0.897423
87366	lebestm kid tablet md	0.865241
124058	rexipra lite	0.863248
91277	mucimega effervescent tablet orange	0.857558
37956	dynagliptm forte tablet sr	0.857369

**Figure 4.0.1:** Cosine Similarity Results for Identification

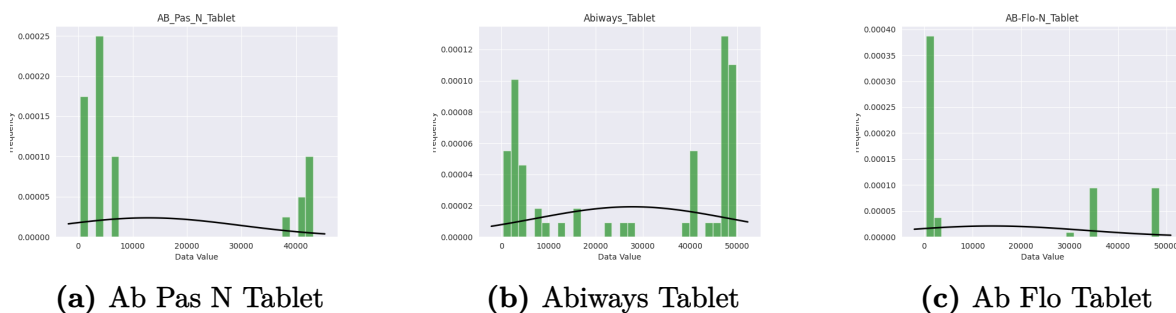
Now, this meta information is stored as shown in Appendix A. Now this meta information has five features: text, color moments, texture, shape, and pattern, and has associated values surrounding the same. In comparison, we get a metric as discussed in Chapter 3.

Now on plotting these values, we get varying results. For instance in Figure 4.0.2 we see that the consistency is not a desired outcome. It is clear from the plots that the values exhibit an inconsistent pattern and have a large range of values that consists of various outliers.



**Figure 4.0.2:** Consistency Analysis

Again, we observe from figure 4.0.3 that we are not able to determine the normality of the values and are sparsely distributed throughout.



**Figure 4.0.3:** Consistency Analysis

Next, the next result storing the data in the IPFS file system which stores out. Now, this meta information can be leveraged again by changing information by counterfeiters. To

prevent this decentralization is necessary. IPFS uses content addressing which means, even though in the worst case the counterfeiters get the address of these files, changing the content will lead to a change in its location. Decentralization, speed, and end-to-end encryption are other salient features of the IPFS file system.

On checking for normal distribution using `shapiro` from the SciPy library, we see that there is no normal distribution. Hence we calculate the upper threshold of the interquartile range as the measure of the threshold. We moreover implement the functions using FastAPI to call functions and verify for identification and threshold.

## 4.1 | Project Outcomes

The following are the Project outcomes:

- Counterfeit recognizing system: The approach developed out of this project is capable of detecting and analyzing counterfeit products. Unlike most machine learning models, this approach is computationally efficient and highly scalable, making it a promising solution for large-scale applications.
- Nascent Approach: The milestones achieved in this project have led to the development of a new approach that addresses the common problem of limited datasets in machine learning projects. This novel cross-analysis methodology is computationally efficient and can be applied in situations where traditional machine-learning methods fall short. Furthermore, this project has distinguished itself from existing projects, as evidenced in Section 2.
- Dataset: While the obtained dataset has some limitations, it is currently the only one of its kind in India, highlighting a need for more similar datasets to increase public awareness and support further research in this area.
- Data Analysis Methodology: The project was able to create a robust and user-friendly function that can be easily replicated to conduct in-depth image analysis. This function will provide a comprehensive set of information and enable researchers to gain deeper insights into their data.

## 4.2 | My Contributions to the Project

I collaborated with Dr. Kundu to generate and discuss the initial ideas for the project. Through these ideation sessions, we brainstormed and refined the project's objectives, scope, and deliverables.

I was solely responsible for the implementation of the project, including the planning, execution, and completion of all project tasks. My contributions to the project included conducting research, collecting and analyzing data, creating project deliverables, and presenting the project findings. Overall, Dr. Kundu and I worked together to conceptualize the project, but I took ownership of its execution and delivery.

## 4.3 | Learning Outcomes

Some of the learning outcomes of this project are:

1. **Proficiency in NLP Techniques:** The project involves using natural language processing techniques to analyze metadata associated with medicines. Got proficiency in NLP concepts such as Name Entity recognition, tokenization, and Topic modeling.
2. **Familiarity with Image Processing Techniques:** The project involves using various image processing techniques to extract metadata from medicine images. I was able gain familiarity with techniques such as image segmentation, feature extraction, and image recognition.
3. **Knowledge of Machine Learning Algorithms:** The project required using machine learning algorithms to classify medicines as genuine or counterfeit based on patterns identified in the metadata. However, due to dataset that did not work out, still explore Yolov5 for multi class classification
4. **Experience with Scraping Data:** The project involves scraping data from various sources to extract a list of medicines and their associated images. I was able to get hands-on experience with web scraping techniques and tools such as BeautifulSoup and Scrapy.
5. **Understanding of Data Security:** The project involves storing metadata associated with medicines in a secure and encrypted format in an IPFS file system. Understood the concepts of data security concepts such as encryption, digital signatures, and blockchain.

## 4.4 | Research Outcomes

The research of identifying counterfeits is still at an early stage. Using no extra equipment such as a spectrometer for authentication of these medicines is an ideal situation for the everyday person. Following are the research outcomes which this project was able to achieve:

- **Counterfeit Medicine Detection Technique:** Cross-Analysis is a low-computing and promising approach for the identification and categorizing it as fake or real based solely based on the appearance of the medicine. This accuracy can significantly improve if a better dataset is built using similar background and photographic techniques.
- **Developing Cross Analysis Technique:** As discussed in Chapter 3 we understand the dimensionality and the behaviour of various features and develop a formula where we imagine the metrics received via Jaccardian similarity and Euclidean distance in a 5-Dimensional space to calculate a metric.
- **Inconsistent Dataset approach:** The dataset limited the project from using machine learning techniques which, as expected, resulted in much lower accuracies leading to taking a nascent approach and finding a mathematical threshold to achieve a rough estimate of identification as authentic or fake medicine.

- **Scalability:** Along with having a low computation system, this system is extensively scalable. Feature extraction and storing of JSON files requires extremely less space and time. The computations have a maximum usage of  $O(N)$ . Currently, the analysis is done on 115 medicine classes, however, expanding this to several of thousands of medicinal classes should be straightforward.
- **Dataset:** As discussed several times, there is a lack of dataset and collection of this dataset could be a good starting point for future work and refining of the dataset.

Now, these research outcomes could be the basis of future work on this analysis of images. Some of them may include Optimising mathematical formulas for weighted features which might variate from medicine to medicine. Along with this, one could figure out a better way of calculating the threshold other than the upper threshold of the inter-quartile value for representing such dispersed data. Another future aspect could be developing a score such as **F1-score** to get a visualization of the accuracy of the cross-analysis of features.

## 4.5 | Real World Applications

Some of the real-world applications of this system are:

1. **Pharmaceutical Companies:** Pharmaceutical companies can use this system to identify counterfeit versions of their own products, which can help them take appropriate legal action against counterfeiters and protect their brand reputation. They can also define the metadata information using digital signatures for better identification of counterfeiters.
2. **Regulatory Bodies:** Regulatory bodies such as the FDA (Food and Drug Administration) can use this system to monitor the market for counterfeit medicines and take appropriate action to remove them from circulation, thereby protecting public health.
3. **Healthcare Providers:** Healthcare providers can use this system to identify counterfeit medicines before administering them to patients, which can help prevent adverse reactions or complications due to fake drugs (which potentially can have adulterants present).
4. **Patients:** Patients can use this system to check the authenticity of the medicines they are prescribed or purchased, which can help them avoid counterfeit drugs that may be ineffective or harmful.
5. **Law Enforcement Agencies:** Law enforcement agencies can use this system to identify and track down counterfeiters involved in producing and distributing fake drugs, which can help prevent the spread of counterfeit medicines and protect public health.
6. **Identification Methodology:** The method implemented in this project can be replicated in real-world for identification in case there is a lack of dataset for training.
7. **Security and Integrity of Metadata:** The project explores encryption techniques and the IPFS file system to store and secure the metadata associated with medicines. The research can explore the effectiveness of encryption algorithms and the robustness of the IPFS file system in protecting the integrity and privacy of the metadata.

8. Integration of Cloud Infrastructure: The project incorporates Digital Ocean Droplets and Spaces for hosting and storing data. The research focuses on the performance, scalability, and cost-effectiveness of utilizing cloud infrastructure for counterfeit medicine detection systems.

Overall, the system can have a significant impact on public health and safety by enabling the identification and removal of counterfeit medicines from circulation, thereby preventing harm to individuals and healthcare systems.



# Chapter 5

## Conclusion

In conclusion, the proposed system offers an efficient and scalable approach to classify images as counterfeit medicines. The system is built on concepts from Natural Language Processing and Cross analysis of information obtained from information.

Now along the way, various techniques involving machine learning and deep learning were utilized which failed due to the inconsistencies in the dataset. This dataset, however, with more information could make it significantly useful and could be efficiently used in these techniques. One such way would be access to the fake images which are not publicly available. Such images could provide a better understanding of a machine-learning model making it easier to classify the dataset.

Overall the proposed system offers an efficient and cost-effective solution to these problems. This system could be highly scalable. Moreover, the system can have higher accuracy if a central high-quality dataset (similar images from a single source -including the resolution) is fed to it which will have more accurate information with less Jaccardian and Euclidian distances. By utilizing this minimum viable product, a helpful system for the public can be built.

# Bibliography

- [1] B. J. Ferdosi, M. A. Sakib, M. S. Islam, and J. Dhar, "Identifying counterfeit medicine in bangladesh using deep learning," in *Human Centred Intelligent Systems: Proceedings of KES-HCIS 2021 Conference*, pp. 46–55, Springer, 2021.
- [2] H.-W. Ting, S.-L. Chung, C.-F. Chen, H.-Y. Chiu, and Y.-W. Hsieh, "A drug identification model developed using deep learning technologies: experience of a medical center in taiwan," *BMC health services research*, vol. 20, no. 1, pp. 1–9, 2020.
- [3] M. Alsallal, M. S. Sharif, B. Al-Ghzawi, and S. M. M. al Mutoki, "A machine learning technique to detect counterfeit medicine based on x-ray fluorescence analyser," in *2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, pp. 118–122, IEEE, 2018.
- [4] A. K. Mishra and M. H. Essop, "Low-cost spectrogram based counterfeit medicine detection," *arXiv preprint arXiv:1904.07152*, 2019.
- [5] K. Abbas, M. Afaq, T. Ahmed Khan, and W.-C. Song, "A blockchain and machine learning-based drug supply chain management and recommendation system for smart pharmaceutical industry," *Electronics*, vol. 9, no. 5, p. 852, 2020.
- [6] S. R. Shinde, K. Bhavsar, S. Kimbahune, S. Khandelwal, A. Ghose, and A. Pal, "Detection of counterfeit medicines using hyperspectral sensing," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 6155–6158, IEEE, 2020.

# Appendix

## Appendix A

```
1  {
2    "Text": { "Word List": ["paracetin", "acceptsp"] },
3    "Color Moments": {
4      "Colour Features": [
5        1.0, 110.39896456507962, 16528.484806076085, 114.30817309996748,
6        12446.342583695696, 17624.072712679976, 1.0, 110.67421810535434,
7        16562.478133236757, 114.38506626520892, 12483.169998915238,
8        17603.566928310596, 1.0, 109.8632574367846, 16429.417306259704,
9        114.19925650376972, 12371.583728220541, 17625.1121283673
10     ]
11   },
12   "Texture": {
13     "Contrast": 76.36967488789237,
14     "Dissimilarity": 4.506606342088405,
15     "Homogeneity": 0.483405099902864,
16     "Energy": 0.28716770423826343,
17     "Coorelation": 0.983637746559528
18   },
19   "Shape": {
20     "Mean Area": 49729.0,
21     "Std Area": 0.0,
22     "Mean Perimeter": 892.0,
23     "Std Perimeter": 0.0,
24     "Mean Aspect Ratio": 1.0,
25     "Std Aspect Ratio": 0.0,
26     "Mean Centroid X": 112.0,
27     "Std Centroid X": 0.0,
28     "Mean Centroid Y": 112.0,
29     "Std Centroid Y": 0.0
30   },
31   "Pattern": { "Entropy": 5.898889541625977 }
32 }
```

JSON file

## Appendix B


### The Importance of IPFS file System:

The IPFS (InterPlanetary File System) system plays an important role in the project for storing and retrieving the metadata associated with medicines. Here are some key reasons why IPFS is important in this context:

- **Decentralized Storage:** IPFS is a decentralized file system that distributes data across a network of participating nodes. In the context of counterfeit medicine detection, this ensures that the metadata associated with medicines is not stored in a single centralized location, making it more resilient against single points of failure and potential tampering.
- **Immutable Content Addressing:** IPFS uses content addressing to uniquely identify data based on its content, rather than its location or other metadata. Each piece of data is given a unique cryptographic hash based on its content, ensuring the integrity of the stored metadata. This is crucial for ensuring that the metadata associated with medicines remains unchanged and tamper-proof.
- **Data Integrity and Verification:** IPFS uses a versioning system that allows previous versions of data to be referenced and retrieved. This ensures that the history of the metadata associated with medicines is preserved, enabling verification and auditing of changes over time. It helps maintain a transparent and trustworthy record of the metadata associated with each medicine.
- **Efficient Data Distribution:** IPFS utilizes a distributed hash table (DHT) to efficiently locate and retrieve data from the network. This enables fast and reliable access to the metadata associated with medicines, regardless of the geographical location of the data and the user. It ensures that the system can handle a large volume of data and provides scalable storage for the project.
- **Security and Privacy:** IPFS provides cryptographic mechanisms to ensure the security and privacy of the stored data. The metadata associated with medicines can be encrypted using algorithms such as RSA before being stored on IPFS, ensuring that sensitive information remains confidential and protected.

By leveraging the IPFS system, the project can benefit from decentralized and secure storage of the metadata associated with medicines. It enhances data integrity, availability, and scalability while providing a robust and reliable infrastructure for storing and retrieving the essential information required for counterfeit medicine detection.

The figure below shows storing metadata extracted from images and pinned in the IPFS file system.



IPFS

STATUS

FILES

EXPLORE

PEERS

SETTINGS

Revision 2191305  
See the code  
Report a bug


Browse
?
+

Files

1MiB FILES

9MiB ALL BLOCKS

+ Import

Name ↑	Pin Status	Size
<div>metaanalysis</div> <div>QmQP47Vh5Re2NFe558rCBt5f68NJs4Yobe1N3215gkxUtv</div>		1 MiB ...

## IPFS File System